基于密度聚类方法在文本挖掘中的应用研究

茅剑[1], 吴顺祥[2]
(1. 厦门大学自动化系 福建 厦门 361005 2. 厦门大学系统与控制研究中心 福建 厦门 361005)

【摘 要】: 文本聚类是文本挖掘的重要组成部分。本文详细分析了文本聚类的过程,并给出了一个文本聚类模型。分析比较各类聚类算法之后,着重研究了一个基于密度的聚类算法,以及它在文本挖掘中的具体应用。

【关键词】: 文本挖掘 聚类 密度 DBSCAN

1. 概述

文本挖掘作为数据挖掘的一个分支, 它把文本型信息源作为分析的对象,利用定量计算和定性分析的方法,从中寻找信息的结构、模型、模式等各种隐含的知识。文本挖掘又是一项综合技术,涉及数据挖掘、计算机语言学、信息检索、自然语言管理、知识管理等诸多领域。

文本聚类是文本挖掘中的一项重要技术。它通过对文本内容的分析,将原始文本集合分成若干个簇,要求簇内文本内容的相似性尽可能大, 而簇之间的相似度尽可能小。文本聚类可广泛应用于文本挖掘与信息检索的不同方面。文本聚类在大规模文本集的组织与浏览、文本集层次归类的自动生成等方面都具有重要的应用价值。本文论述的文本聚类模型如图 1 所示:

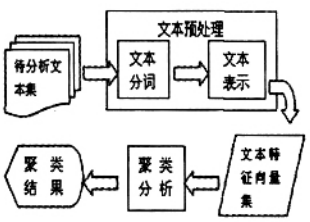

图 1

2. 文本预处理

本文讨论汉语言文本的聚类问题。在进行聚类分析之前,需要对文本集进行预处理, 将文本格式转化成聚类算法能够处理的数据格式。需要指出的是,文本预处理的应用不仅限于文本聚类,它同样是文本挖掘的第一个步骤,对文本挖掘效果的影响至关重要。

2.1 文本分词

目前,在信息处理方向上的基本思想是以向量来表示文本,那么选取什么作为特征项呢,一般可以选择字、词或词组,根据实验结果,普遍认为选取词作为特征项要优于字和词组。因此,文本预处理的第一步工作就是文本分词, 它是进行文本表示的准备工作。为了将一篇文档切分为一个个独立的中文单词,本文采用了一种逆向最大匹配算法。算法基本思想是:从文档的最后一字开始,逆向抽取最大长度的字符串,然后在词典中循环匹配可以独立成词的子字符串,逆向搜索文档直至无词可取。

分词中要注意的是,为了更好的体现出文本的语义内容,需要在分词过程中去除停用词(不宜作为文本特征的词,如介词、叹词等)和合并同义词(将意思相近的词用一个词替代,如"微机"可用"计算机"替换)。停用词典和同义词典需要事先定义。对于之后的文本表示来说,这实际上是一个特征值约减的降维过程。

2.2 文本表示

文本的表示主要采用向量空间模型(VSM)。向量空间模型的基本思想是以一个规范化的特征向量来表示文本:V (d)=(t1, W1; … ti,Wi; … tn,Wn)。其中 ti 为第 i 个特征项, Wi 为第 i 个特征项的权重。在之前文本分词的基础上,把切分后的词作为向量的特征项 ti,将统计过的词频计算后作为向量的权重 Wi 来表示文本。词频分为绝对词频和相对词频,绝对词频,即使用词在文本中出现的频率表示文本;相对词频为归一化的词频。

文本中词空间维度很高, 不同的词对文本内容的贡献是不等的。因此需要度量词在文本中的权重,只有大于一定权重阀值的词才能作为表征文本内容的关键词。提取关键词的工作就称为文本特征抽取。文本特征抽取的本质是高维数据的降维技术,即将高维数据通过变换映射到低维空间。降维的主要问题在于,从高维到低维的变换可能掩盖数据原有的信息。这样原先在高维空间存在明显差异或特征的类别在低维空间内会混杂在一起难以区分。因此,寻找合适的映射就成了文本特征抽取的关键。

在文本中的词的加权体系中用某一权重值取代表示该词是否出现的{0,1}二进制布尔表示,通常具有更高的准确性。词在文本中的权重计算方法主要运用 TF- IDF 公式, 目前存在多种 TF- IDF 公式,本文中采用的 TF- IDF 公式:

$$W(t,\vec{d}) = \frac{(1+\log_2 tf(t,\vec{d}))\times \log_2(N/n_t)}{\sqrt{\sum_{t\in\vec{d}}\left[(1+\log_2 tf(t,\vec{d}))\times \log_2(N/n_t)\right]^2}}$$

其中, W(t,d)为词 t 在文本 d 中的权重, 而 tf(t,d)为词 t 在文本 d 中的词频, N 为训练文本的总数, n_t 为训练文本集中出现 t 的文本数,分母为归一化因子。

3. 文本聚类算法

通过文本预处理, 现已得到了可供聚类算法直接处理的标准数据集。接下来要考虑的是,如何对数据集进行有效的分析,得出所需的聚类结果。聚类算法的选择显得至关重要。

3.1 传统的聚类算法及其局限性

聚类分析作为数据挖掘中的一种分析方法, 它可以作为一个单独的工具以发现数据库中数据分布的一些深入的信息,并且概括出每一类的特点, 或者把注意力放在某一个特定的类上以作进一步的分析; 聚类分析也可以作为数据挖掘算法中其他分析算法的一个预处理步骤。目前, 已经提出的聚类算法有很多。

(1) 划分法: 给定一个有 N 个对象的数据集, 构造 K 个分组,每一个分组代表一个簇(K<N)。对于给定的 K,可以给出一个初始的划分方法, 以后通过迭代重定位将每个对象归入合适的簇中,从而完成聚类划分。如 K- MEANS 算法。划分法的缺陷在于需要事先确定聚类划分的个数 K。而聚类算法的目的正是要将未知数据集做出合适的划分。可见,这种要得到聚类结果必须先确认聚簇个数的做法显然会影响到分析结果的客观性。

(2) 层次法:对给定的数据集进行层次的分解,直到某种条件满足为止。层次法又分为自底向上的凝聚法和自顶向下的分裂法。其基本算法均是迭代分层至某个终止条件。层次法的缺陷在于,一个步骤(合并或分裂)完成,就不能被撤销。因此需要和其他算法结合来改进聚类结果。

(3) 基于网格的方法:将数据空间划分成有限个单元的网格结构,所有的处理都是以单个的单元为对象的。从而使得处理速度加快,因为它与目标数据库中记录的个数无关,只与数据空间的单元多少有关。

(4) 基于模型的方法:给每一个聚类假定一个模型,然后去寻找能够很好的满足这个模型的数据集。这样一个模型可能是

基金项目:由厦门大学 985 二期信息创新平台项目和福建省教委科技项目(JA05290)资助

数据点在空间中的密度分布函数或者其它。它的一个潜在的假定就是:目标数据集是由一系列的概率分布所决定的。通常有两种尝试方案:统计的方案和神经网络的方案。

3.2 DBSCAN 算法描述及其实现

DBSCAN 算法是一种基于密度的聚类算法。它利用类的密度连通特性,可以快速发现任意形状的类。其优点在于,它可以发现任意形状的聚类,并且不受"噪声"的干扰

基于密度聚类的关键思想是,聚类中每个核心点在给定半径(ε)的圆内的相邻对象至少必须达到一个数量(MinPts),也就是相邻对象的数量必须超过一个阀值。下面对一个聚类的定义有简短的介绍。

定义一:直接密度可达

在对象集 D 中,关于ε和 MinPts,对象 p 直接密度可达至对象 q 必须满足以下条件:

1) $p \in N_\varepsilon(q)$ ($N_\varepsilon(q)$ 是对象集合 D 包含在 q 的 - 邻域内的子集)

2)$Card(N_\varepsilon(q)) \geqslant MinPts$(Card(N) 表示集合 N 内的对象数量)

条件 2) 称为" 核心对象条件"。如果一个对象 p 满足该条件,则称 p 为一个"核心对象"。其它对象只有相对于核心对象才能称为直接密度可达。

定义二:密度可达

在对象集 D 中,关于 ε 和 MinPts,对象 p 密度可达至对象 q 必须满足以下条件: 存在一个对象链 p1, ..., pn, p1 = q, pn = p,其中 $pi \in D$ 且 pi+1 关于∈和 MinPts,直接密度可达至 pi。

密度可达是直接密度可达的传递。密度可达的两个对象关系通常上是不对称的。也就是说,p 密度可达 q 并不意味着 q 密度可达 p,只有核心对象之间才可以相互密度可达。

定义三:密度相连

在对象集 D 中,关于ε和 MinPts,对象 p 与对象 q 密度相连必须满足以下条件:

存在一个对象 $o \in D$, p 和 q 关于ε和 MinPts, 均密度可达至 o 密度相连的对象是一个对称的关系。

定义四:聚类和噪声

在对象集 D 中,一个关于ε和 MinPts 的聚类 C 是 D 的一个非空子集。它满足以下条件:

1) 极大性: $\forall p,q \in D$: 如果 $p \in C$ 且 q 密度可达 from p 关于ε和 MinPts,而且 $q \in C$.

2) 连通性: $\forall p,q \in C$: p 密度可达 to q 关于 D 中的 ε 和 MinPts 。

不属于任何聚类的对象为噪声。

根据以上定义,使用 DBSCAN 算法发现的一个聚类,实际相当于数据集中以一个核心对象为起点, 密度可达的所有对象的集合。而密度可达对象的获取实际上是直接密度可达对象获取过程的重复。

DBSCAN 算法的处理步骤如下:

(1) 检查数据库中每个对象的 - 邻域,如果一个对象 p 的 - 邻域 $N_\varepsilon(p)$内包含的对象数目大于 MinPts,就创建一个包含 $N_\varepsilon(p)$中所有对象的新聚类 C。

(2) 检查 C 中还未处理过的对象 q, 如果 q 的 - 邻域 $N_\varepsilon(q)$ 内包含的对象数目大于 MinPts,就把 $N_\varepsilon(q)$中未包含于聚类 C 的对象加入聚类 C 中。

(3) 对这些对象做同样的处理。重复 (1) (2),直到没有新的对象可以加入聚类 C。

给定数据集 DataSet,聚类半径 ε,最小对象数量 MinPts,有 DBSCAN 算法代码实现:

```
DBSCAN(DataSet, ε Minpts)
{ClusterId=nextId(NOISE);
    for (int i=1; i<DataSet. Size; i++)
    {       Point=DataSet.get(i);
        if (Point.Clid==UNCLASSIFIED)
            if (ExpandCluster(DataSet, Point, ClusterId, εMinpts))
                ClusterId=nextId(ClusterId);
     }
}
```

4. 聚类实验及结果分析

本文从网上采集了 3000 篇文档,作为初始文本集。文档表现为:文档内容复杂,按照不同的网页主题,共采集了 10 类文档;文档大小差别悬殊,去噪后的文档字节数从 83Byte 到 38KB 不等。

按照本文中所述的步骤进行实验。结果表明,如果各文档主题之间区别明显且文档分布稠密,实验效果很好,聚类结果和文档的类别基本相符,且不受噪声干扰。但是如果各文档的主题接近或文档内容主题不鲜明, 则有可能出现将几个聚类簇合并的现象。本文分析认为,在整个文本聚类过程中,文档的特征表示对聚类结果的影响是举足轻重的。

此外, 本文还用了 K- MEANS 算法进行了比对运算实验, 实验结果如图 2 所示。由此可见在大数据量的情况下,DBSCAN 算法时间复杂度低的优势可以得到很好的体现。这说明 DBSCAN 算法不失为一个高效、快速的聚类算法, 它在文本聚类中可以得到很好的应用。

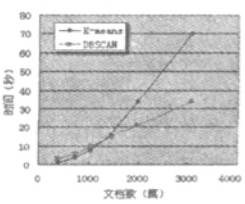

图 2

4. 总结

文本聚类尤其是 Web 文本聚类,在互联网飞速发展的今天有着广阔的应用前景。基于密度的聚类算法以其快速高效,抗干扰性强的优点,可以很好地适应海量数据文本挖掘的要求。

参考文献:

1. Alsabti K,Ranka S, Singh V.An Efficient K- means Clustering Algorithm [C]. Proceedings of the First Workshop on High Performance DateMining, IPPS- 98, Orlando, Florida,USA,1998

2. ESTER M, HANS2PETER KR IEGEL, SANDER J, et al. A density2based algorithm for discovering clusters in large spatial databaseswith noise [A]. Proceeding the 2nd international conference onKnowledge discovery and data mining (KDD) [C]. Portland, 1996.226 - 231.

3. Han J, Kamber M. Data Mining: Concepts and Techniques. SimonFraser University,2000

4. Mihael Ankerst, Markus M. Breunig, Hans- Peter Kriegel, J?rg Sander: OPTICS: Ordering Points To Identify the Clustering Structure Proc. ACM SIGMOD'99 Int. Conf. on Management of Data, Philadelphia PA, 1999.

5. 苏新宁等著. 数据挖掘理论与技术. 北京:科学技术文献出版社,2003

(上接第 14 页)

参考文献:

1. M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, Image Inpainting, SIGGRAPH 2000, pp. 417- 424.

2. Chan, T., Shen, J. Mathematical Models for Local Deterministic Inpaintings. UCLA CAM TR 00- 11, March 2000.

3. Chan, T., Shen, J. Non- Texture Inpainting by Curvature- Driven Diffusions (CCD). UCLA CAM TR 00- 35, Sept. 2000.

4. M. M. Oliveria, B. Bowen, R. Mckenna, and Y. Chang, Fast Digital Image Inpainting, Proceedings of the International Conference on Visualization, Imaging and Image Processing (VIIP 2001), pp. 261- 266, 2001.

5. 丁雯, 一类非线性扩散问题及其在图像修复中的应用, 上海交通大学学报, 2004, 38(1):153- 156.

6. 邵肖伟,刘政凯,宋璧, 一种基于 TV 模型的自适应图像修复方法, 电路与系统学报, 9(2): 113- 117,2004

7. 周廷方,汤锋,王进,王章野,彭群生,基于径向基函数的图像修复技术, 中国图象图形学报, 9(10):1190- 1196, 2004.